APPLIED MSC IN DATA SCIENCE & ARTIFICIAL INTELLIGENCE

MACHINE LEARNING WITH PYTHON LABS

STROKE PREDICTION PROJECT

Submitted by: Maryam Nasir || Professor: Hanna Abi Akl

DSTI SPOC 2025 maryam.nasir@edu.dsti.institute

Contents

Introduction

This document presents the interpretation of results from a data analysis project aimed at predicting stroke, along with the steps taken throughout the process. It was developed as part of the Machine Learning with Python Labs course at DSTI. The report is designed to be read alongside the accompanying Jupyter notebook in DSTI-ML-PY-STROKES under `notebooks\StrokePredictionEDA.ipynb`. It provides a structured overview of the dataset, its processing and cleaning, the exploratory data analysis (EDA), feature selection, results obtained, key limitations, and final conclusions.

Data Description

The data provided is used to predict whether a patient is likely to get a stroke based on input parameters like gender, age, various diseases and smoking status. Below is the information we have regarding the dataset attributes:

1) **id**: unique patient identifier
2) **gender**: "Male", "Female" or "Other"
3) **age**: age of the patient
4) **hypertension**: 0 (if the patient doesn't have hypertension) or 1 (if the patient has hypertension)
5) **heart_disease**: 0 (if the patient doesn't have a heart disease) or 1 (if the patient has a heart disease)
6) **ever_married**: "No" or "Yes"
7) **work_type**: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
8) **Residence_type**: "Rural" or "Urban"
9) **avg_glucose_level**: average glucose level in the blood
10) **bmi**: body mass index
11) **smoking_status**: "formerly smoked", "never smoked", "smokes" or "Unknown" (in this case the information for the patient is not available)
12) **stroke**: 1 (if the patient had a stroke) or 0 (if the patient didn't have a stroke)

```
Index: 5110 entries, 9046 to 44679
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             5110 non-null   object
 1   age                5110 non-null   float64
 2   hypertension       5110 non-null   int64
 3   heart_disease      5110 non-null   int64
 4   ever_married       5110 non-null   object
 5   work_type          5110 non-null   object
 6   Residence_type     5110 non-null   object
 7   avg_glucose_level  5110 non-null   float64
 8   bmi                4909 non-null   float64
 9   smoking_status     5110 non-null   object
 10  stroke             5110 non-null   int64
dtypes: float64(3), int64(3), object(5)
memory usage: 479.1+ KB
```

```
df.isna().sum()
```

```
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

Figure 1: Showing concise summary of dataset and missing values

The dataset has 5110 entries with missing values found only in the `bmi` column, where 201 records are incomplete.

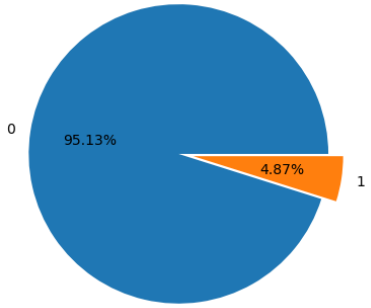


Figure 2: Pie chart showing stroke data ratio

We also discover that the data is highly unbalanced and is not in favor of our target class, i.e., patients with stroke. Roughly **95% of patients recorded in this dataset do not have stroke**.

Here, the plot shows the mean values of the numerical features grouped by patients with stroke and without stroke. On average, most features are higher for stroke patients, with the difference particularly pronounced in age and average glucose levels, suggesting that these factors may have a stronger influence on stroke likelihood.

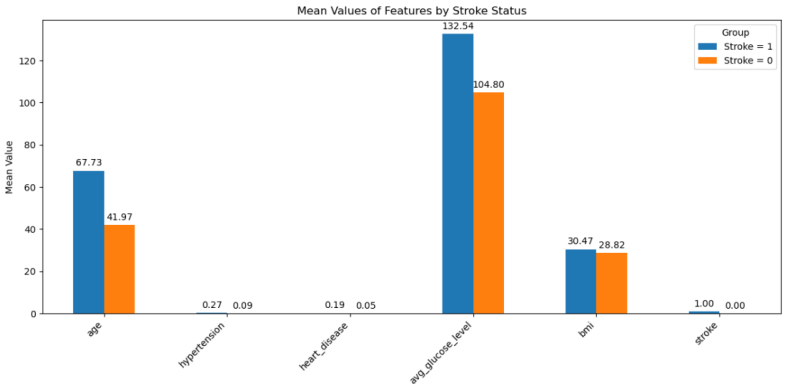


Figure 3: Showing mean value of stoke data distributed across the numerical fields

Data Processing and Cleaning

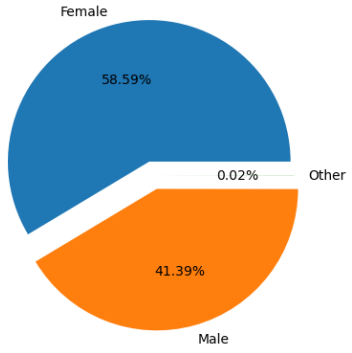


Figure 4: Pie chart showing gender ratio

The dataset has only 1 record (~0.02% of 5110) with gender = “Other”. This patient has no hypertension, heart disease, or stroke history, though their glucose level is above average and they are a former smoker. Since they are not in the target class (stroke = 1), the row adds no predictive value. Instead, it introduces a third gender category, which increases model complexity and overfitting risk, while being statistically insignificant. As such, we make the decision to drop gender = “Other”.

To address the 201 missing BMI values, we used an imputation strategy to fill based on gender and age rather than dropping these as 40 patients have a history of stroke. This is roughly 16% of Strokes data available to us. Since BMI is derived from a patient’s mass and height (data not available in this dataset), we instead calculated the average BMI for groups defined by gender and age. After ensuring there were no missing values in these reference groups, we assigned the corresponding group average to patients with missing BMI, providing a more contextually accurate estimate than a simple overall mean.

```
gender  age
Female  0.08     14.100000
        0.32     17.266667
        0.40     17.400000
        0.48     16.100000
        0.56     18.300000
                    ...
Male    78.00    27.247222
        79.00    27.681818
        80.00    29.210714
        81.00    27.677273
        82.00    27.943478
Name: bmi, Length: 205, dtype: float64
```

Figure 5: Showing the average BMI grouped by gender and age

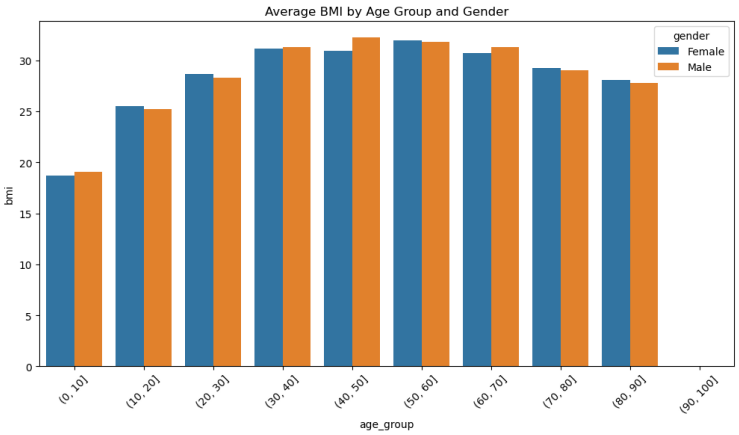


Figure 7: Showing the distribution of BMI by gender across the age groups

```
State of missing values before processing
201
State of missing values after processing
0
```

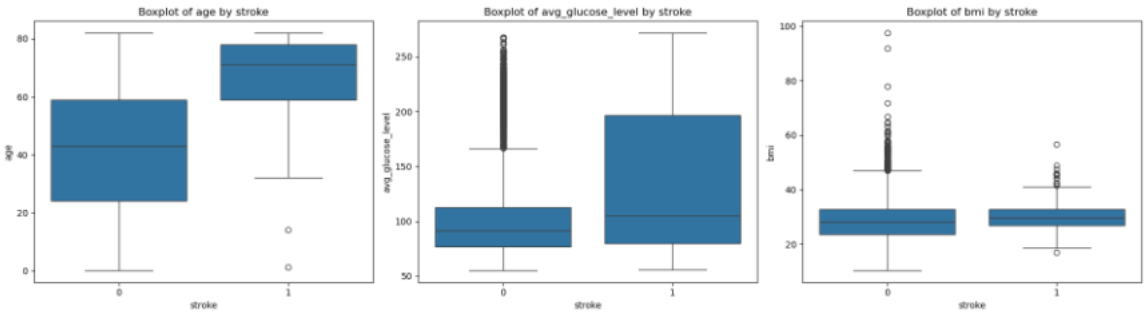
Figure 6: Showing filling of BMI

Outliers (For Trail 2)

For trail 2 after the initial modelling, outliers in age, avg_glucose_level, and bmi were capped separately for stroke=1 and stroke=0 using the 1.5×IQR method. This adjustment was applied after the initial modelling and validation, not because the first results were poor, but to explore whether refining the data could further improve model accuracy. This is further commented on in the validation section of the report.

For each class, lower and upper whiskers were calculated, and values outside these limits were clipped. In age, stroke=1 had 2 extreme values capped on the lower whisker, while stroke=0 had no outliers. Avg_glucose_level had no outliers in stroke=1, but 552 extreme values in stroke=0 were capped above the upper whisker. For bmi, 15 high stroke=1 values and 114 stroke=0 outliers were capped, mostly above the upper whisker. This method preserves the class distribution while limiting the influence of extreme values, ensuring that numerical features remain within meaningful ranges for both classes and supporting more stable model training.

Overall, this preprocessing step ensures that numerical features remain within meaningful ranges, preserves minority-class data, and reduces the potential impact of extreme majority-class values on model performance, supporting more robust and stable predictions.




```
age | Stroke=1 | Lower whisker: 30.50, Upper whisker: 106.50, Outliers count: 2
age | Stroke=0 | Lower whisker: -28.50, Upper whisker: 111.50, Outliers count: 0
avg_glucose_level | Stroke=1 | Lower whisker: -95.59, Upper whisker: 372.09, Outliers count: 0
avg_glucose_level | Stroke=0 | Lower whisker: 23.60, Upper whisker: 166.32, Outliers count: 552
bmi | Stroke=1 | Lower whisker: 18.30, Upper whisker: 41.50, Outliers count: 15
bmi | Stroke=0 | Lower whisker: 9.40, Upper whisker: 47.00, Outliers count: 114
```

Figure 8: Box plots showing the outliers for age, glucose level, and bmi by stroke data

Exploratory Data Analysis

The EDA step continues the understanding of the raw dataset by examining the distribution of features and their relationship with stroke occurrence. This process helps identify patterns, highlight potential risk factors, and guide the feature engineering and modeling phases. We first consider the numerical features: age, average glucose level, and BMI, before turning to the categorical variables.

Numerical Features

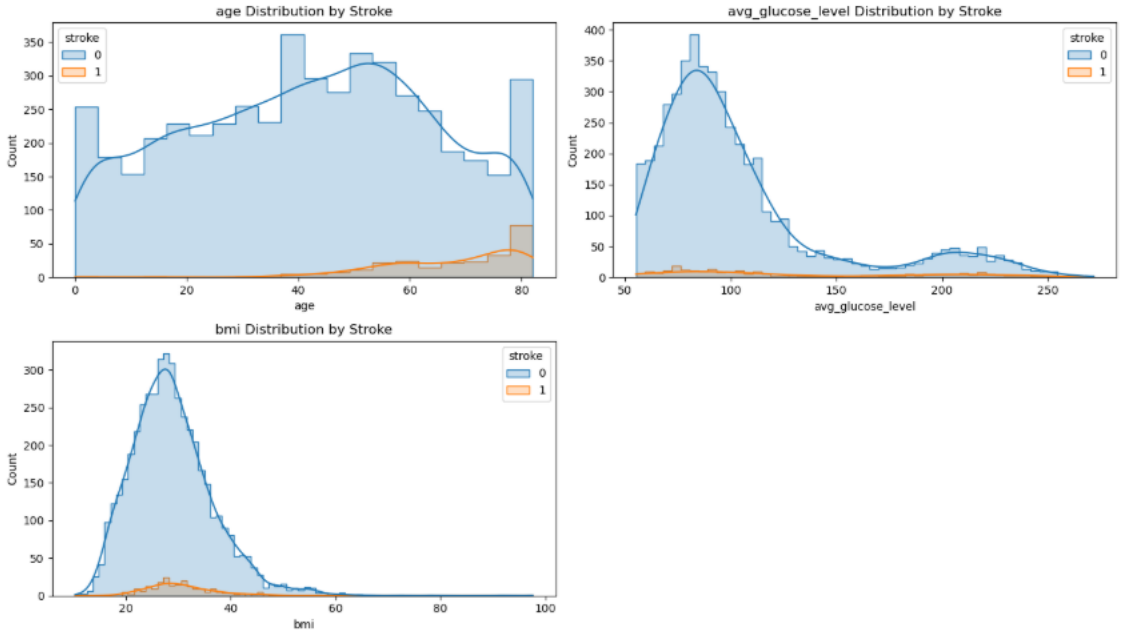


Figure 9: Showing the distribution of age, glucose level and BMI by stroke data

To get a better understanding of these features and their relation to stroke, we can group them and observe their general behavior.

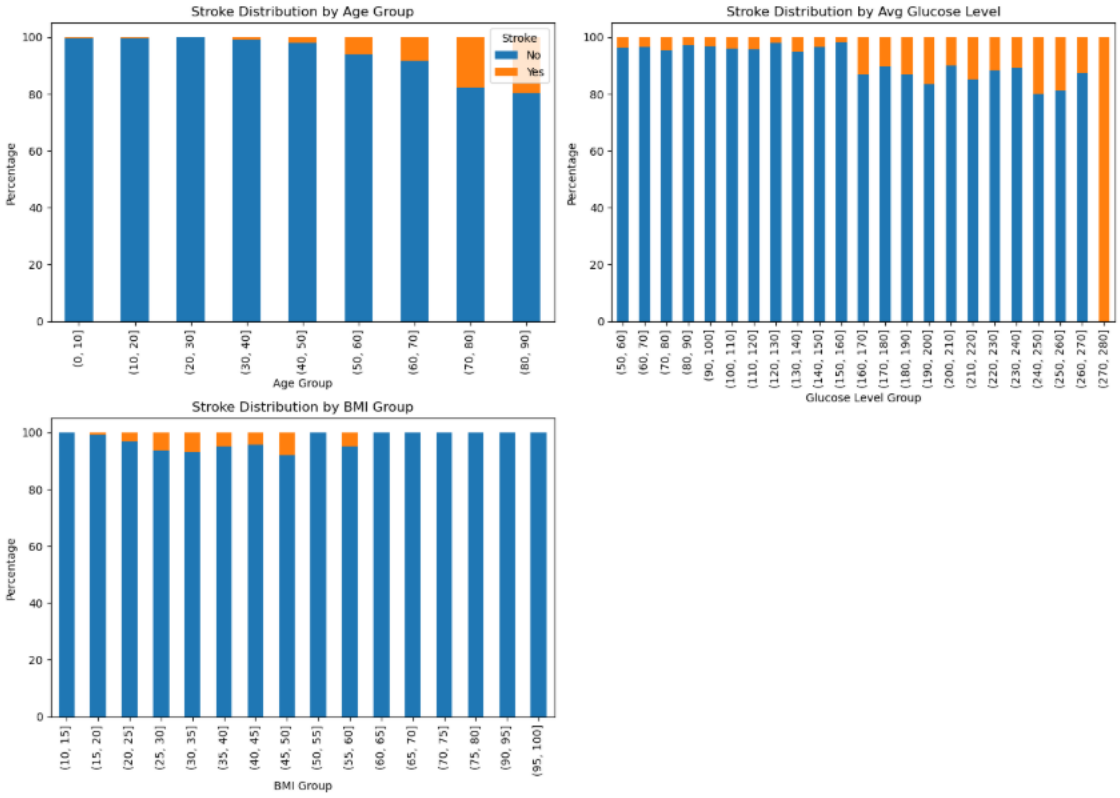


Figure 11: Showing the distribution of grouped age, glucose level and BMI by stroke data

The Age Distribution by Stroke shows a strong skew toward older patients, indicating that stroke likelihood increases significantly with age. Grouping patients into age ranges makes this trend clearer: strokes are rare below 40 but rise in the 50+ population, peaking in the 70-80 and 80-90 groups. This suggests age is one of the strongest predictors of stroke in the dataset.

Looking at the Average Glucose Level Distribution by Stroke, most patients fall between 80-120, where stroke prevalence is relatively low. However, risk rises notably for patients with glucose levels above 150, with the 200+ range showing the highest concentration of stroke cases. This points to a strong association between elevated glucose and stroke risk.

The BMI Distribution by Stroke appears roughly normal across the population, with most patients falling into the overweight and obese categories. Stroke cases cluster most heavily in the 25-35 BMI range, aligning with overweight/obese classifications. Below 20 BMI, strokes are rare, while at very high BMIs (>40) the pattern is less clear due to limited data, with no strong upward trend. Overall, age and glucose levels emerge as strong stroke predictors, while BMI has a moderate but notable influence.

Categorical Features

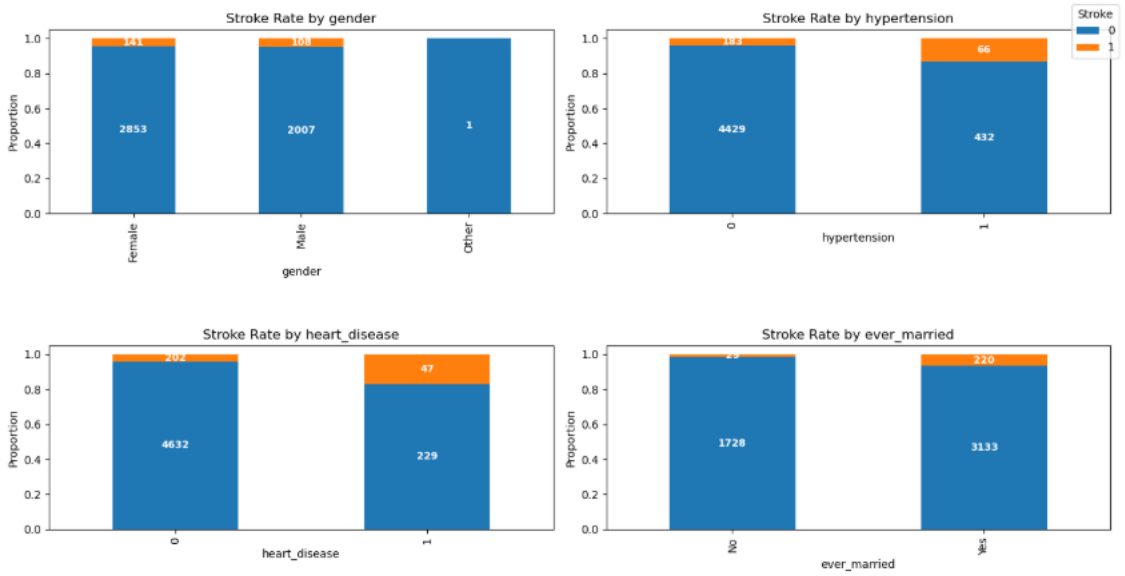


Figure 11: Showing the count of stroke data per category of gender, hypertension, heart disease and marriage status

For categorical features, the Stroke Rate by Gender plot shows that both male and female groups experience strokes at similar rates (~5%). The dataset also contains a single "Other" entry, which is not a stroke patient; as this provides no predictive value, it was excluded during data cleaning.

The Stroke Rate by Hypertension plot reveals that while most patients are not hypertensive (4612 vs. 498), those with hypertension face a much higher prevalence to having stroke (13% vs. 4%). Nonetheless, in absolute terms, more strokes occur in the non-hypertensive group (183 vs. 66) due to its larger size. The Stroke Rate by Heart Disease graph follows a similar pattern: prevalence is higher among patients with heart disease (17% vs. 4%), even though most patients do not have this condition.

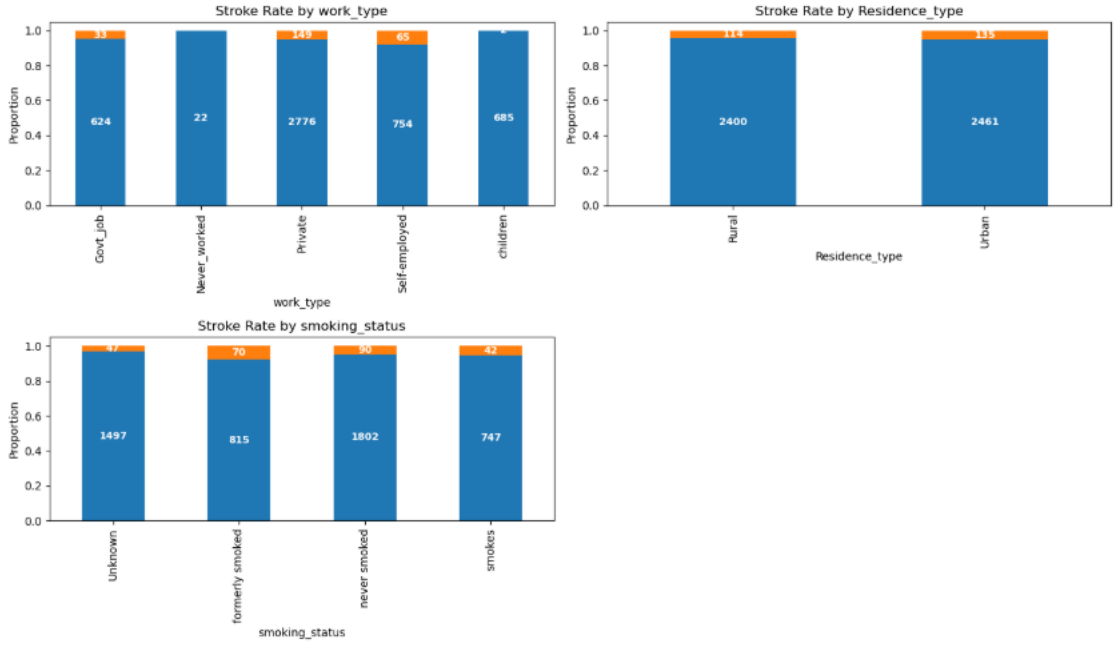


Figure 12: Showing the count of stroke data per category in type of work, residence, and smoking status

In terms of lifestyle and demographics, the Stroke Rate by Marital Status graph shows higher prevalence among married patients compared to unmarried ones. The Stroke Rate by Work Type graph highlights that most strokes occur among those employed in the private sector or self-employed. No strokes were observed in the Never Worked group, and only two occurred among children. The Stroke Rate by Residence Type graph shows patients are fairly evenly split between urban and rural areas, with no meaningful difference in stroke rates. Finally, the Stroke Rate by Smoking Status graph reveals the highest prevalence among formerly smoked, although the majority of patients fall into the never smoked category, where about 5% experienced stroke.

Taken together, these categorical analyses suggest that while stroke cases are relatively rare across all groups, hypertension, heart disease, and smoking history are key risk factors, complementing the strong effects of age and elevated glucose levels observed in the numerical features.

Feature Engineering

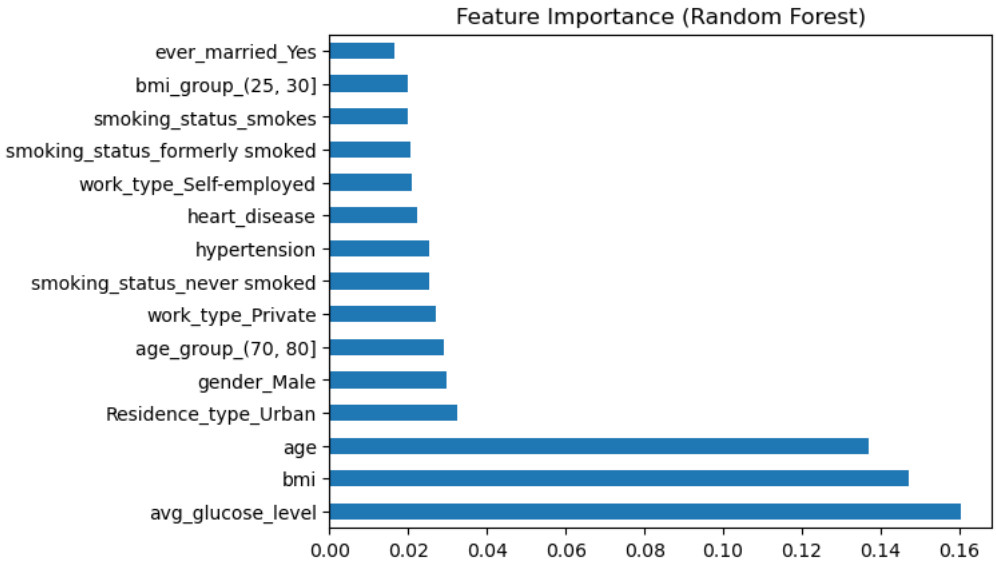


Figure 13: Showing feature importance plot

This plot visualizes the top 15 most important features from a Random Forest classifier to predict strokes. A Random Forest is an ensemble of decision trees. Each tree is trained on a bootstrap sample of the data, and splits are made using random subsets of features. This randomness helps reduce overfitting and improves generalization. A random state of 42 is used to make the results reproducible.

The Random Forest feature importance results show that average glucose level, BMI, and age are the strongest predictors of stroke, contributing far more than any other variables. These three continuous features capture most of the predictive signal, aligning with medical evidence that links elevated glucose, older age, and excess weight to higher stroke risk.

By contrast, demographic and lifestyle factors such as residence type, gender, and specific age groupings (e.g., 70-80 years) contribute moderately, with importance values around 0.03-0.04. Other categorical variables like work type, smoking status, hypertension, heart disease, and marital status, rank among the lowest, each contributing less than 0.03.

Interestingly, features like smoking status, hypertension, and heart disease, which might be expected to have a stronger influence, show less impact in this dataset. This highlights the need to evaluate features objectively and avoid relying on assumptions, as biases in encoding or interpretation can distort model performance.

Overall, the model emphasizes that continuous health indicators (age, glucose, and BMI) dominate stroke prediction, while demographic and lifestyle variables provide supplementary but comparatively limited value. We will test this model later.

Encoding the Categorical Features

Label Encoding is applied to convert the categorical features gender, ever_married, work_type, Residence_type, smoking_status into numeric values required for model training. Each category was mapped to an integer, and a dictionary was created to keep track of the mappings for interpretability later on. Additionally, binary features like hypertension and heart_disease were explicitly labelled with their corresponding values (e.g., 0 = No Hypertension, 1 = Hypertension). This ensures the model can process categorical variables effectively while maintaining clarity on how the encoded values relate to the original categories.

```
{'gender': {1: 'Male', 0: 'Female'},
 'ever_married': {1: 'Yes', 0: 'No'},
 'work_type': {2: 'Private',
  3: 'Self-employed',
  0: 'Govt_job',
  4: 'children',
  1: 'Never_worked'},
 'Residence_type': {1: 'Urban', 0: 'Rural'},
 'smoking_status': {1: 'formerly smoked',
  2: 'never smoked',
  3: 'smokes',
  0: 'Unknown'},
 'hypertension': {0: 'No Hypertension', 1: 'Hypertension'},
 'heart_disease': {0: 'No Heart Disease', 1: 'Heart Disease'}}
```

Figure 14: Showing encoded features

Finally, we get a dataset as seen below for our model:

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	1	67.0	0	1	1	2	1	228.69	36.600000	1	1
51676	0	61.0	0	0	1	3	0	202.21	29.879487	2	1
31112	1	80.0	0	1	1	2	0	105.92	32.500000	2	1
60182	0	49.0	0	0	1	2	1	171.23	34.400000	3	1
1665	0	79.0	1	0	1	3	0	174.12	24.000000	2	1

Figure 15: Showing sample of data to be ingested into out model

Balancing Dataset with SMOTE

To address the significant class imbalance in the dataset where stroke cases (249) are heavily outnumbered by non-stroke cases (4,860), we applied **Synthetic Minority Oversampling Technique (SMOTE)**. SMOTE generates synthetic examples of the minority class rather than simply duplicating existing samples, helping the model learn more generalizable decision boundaries. In this approach, we first reduced the majority class through undersampling to 2,490 non-stroke cases, achieving a manageable ratio of 0.1. We then applied oversampling with SMOTE to increase the minority class to 2,490 stroke cases, resulting in a **balanced dataset of 2,490 samples per class**. This process reduces bias toward the majority class, improves the model's ability to detect stroke cases, and ultimately supports more reliable predictions.

```
stroke
0    4860
1     249
Name: count, dtype: int64
```

$$\text{Sampling Strategy} = \frac{\text{Minority Class Samples}}{\text{Majority Class Samples}}$$

```
Counter({0: 2490, 1: 2490})
```

Figure 16: Showing original stroke data available, sampling strategy used, and the final balanced stroke dataset

Results

First Documented Trail

After EDA and feature engineering, we moved unto the modelling step. To evaluate model performance, the dataset was split into training and testing sets using an 80/20 split. Specifically, 80% of the data was used to train the model, while the remaining 20% was held out for testing. A fixed random seed (random_state = 2) was applied to ensure reproducibility, so that the same split is obtained each time the code is run. For example, included below is a case of feature averages considering this split:

Training Set Averages

```
Training Features:
gender                0.332078
age                  54.888991
hypertension          0.087349
heart_disease         0.053213
ever_married          0.728916
work_type             1.965612
Residence_type        0.407380
avg_glucose_level   118.904712
bmi                  29.642073
smoking_status        1.293926
```

Testing Set Averages

```
Testing Features:
gender                0.305221
age                  55.570445
hypertension          0.082329
heart_disease         0.052209
ever_married          0.722892
work_type             1.941767
Residence_type        0.442771
avg_glucose_level   115.906571
bmi                  29.636653
smoking_status        1.269076
```

Target Class (Stroke) Average

```
Training target:
0.49899598393574296
Testing target:
0.5040160642570282
```

Figure 17: Showing the averages for the training, testing and target values

To evaluate predictive performance, three supervised classification models were implemented: **Logistic Regression, Decision Tree, and Random Forest.** Logistic Regression served as a baseline due to its simplicity, efficiency, and interpretability, estimating stroke probability based on input features. A Decision Tree was used to capture non-linear relationships and feature interactions, offering intuitive, easy-to-visualize decision rules, though it can overfit when used alone. To address this, Random Forest, an ensemble of multiple trees trained on random data subsets, was applied, reducing variance and improving generalization, making it suitable for complex, noisy datasets requiring strong predictive performance.

Logistic Regression

```
---Logistic Regression Results---
Confusion Matrix:
 [[376 118]
 [ 96 406]]

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.76      0.78       494
           1       0.77      0.81      0.79       502

    accuracy                           0.79       996
   macro avg       0.79      0.78      0.78       996
weighted avg       0.79      0.79      0.78       996


ROC AUC: 0.8592391567333904
```

Decision Tree

```
---Decision Tree Results---
Confusion Matrix:
 [[418  76]
 [ 60 442]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.85      0.86       494
           1       0.85      0.88      0.87       502

    accuracy                           0.86       996
   macro avg       0.86      0.86      0.86       996
weighted avg       0.86      0.86      0.86       996


ROC AUC: 0.8633159669016244
```

Random Forest

```
---Random Forest Results---
Confusion Matrix:
 [[434  60]
 [ 35 467]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.88      0.90       494
           1       0.89      0.93      0.91       502

    accuracy                           0.90       996
   macro avg       0.91      0.90      0.90       996
weighted avg       0.91      0.90      0.90       996


ROC AUC: 0.9670447763601464
```

Figure 18: Showing the confusion matrix, classification report and ROC AUC values of the Logistic Regression, Decision Tree and Random Forest Models

Second Trail Removing Outliers

For this Trail, the same models were used, and the dataset was also split 80/20 for training and testing. The only difference lies in the removal of outliers in the numerical data. This is explained earlier in the report under the data cleaning section.

```
---Random Forest Results From Trail 2---
Confusion Matrix:
 [[465  29]
 [ 30 472]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94       494
           1       0.94      0.94      0.94       502

    accuracy                           0.94       996
   macro avg       0.94      0.94      0.94       996
weighted avg       0.94      0.94      0.94       996


ROC AUC: 0.9877272287368744
```

Figure 19: Showing results for the second trail when removing outliers applied to Random Forest Model

Evaluation

In the first trial, Random Forest delivered the strongest performance with 90% accuracy and a ROC AUC of 0.96, clearly outperforming Logistic Regression (79% accuracy, ROC AUC 0.85) and Decision Tree (86% accuracy, ROC AUC 0.86). Random Forest not only had the highest number of correct predictions (434 true negatives and 467 true positives in this instance) but also the lowest misclassifications. Logistic Regression, while useful for its interpretability, produced a high number of false negatives (96 cases), which is particularly concerning in stroke prediction. Decision Trees offered simple decision rules but did not match Random Forest's robustness.

In the second trial, after capping outliers with the 1.5×IQR method, model performance improved further. The best performing model from trail 1 was chosen for further exploration. The Random Forest rose to 94% accuracy and a ROC AUC of 0.98, reducing misclassifications from 95 to only 59 cases and demonstrating exceptional balance between precision and recall.

```
Cross-validation ROC AUC scores: [0.93348898 0.98082692 0.98205472 0.9868611  0.98235311]
Mean ROC AUC: 0.9731169658553893
```

```
Cross-validation ROC AUC scores: [0.9680489  0.98860905 0.99102434 0.9907562  0.99201223]
Mean ROC AUC: 0.9860901437073595
```

Figure 20: Showing the cross validation for k=5 and mean ROC AUC values from the first and second trail of the random forest model

Overall, Random Forest consistently outperformed the other models across both trials, and the outlier treatment further strengthened its reliability. Its superior ability to correctly identify stroke cases (high recall) while minimizing false negatives underscores its value in medical prediction tasks, where the cost of missed diagnoses is critical.

Further validation on the Random Forest model achieved almost perfect performance on the training set, which indicates that it fits the training data almost completely. However, this does not necessarily imply overfitting if the model also generalizes well. To assess this, 5-fold cross-validation was conducted, yielding ROC AUC scores between 0.93 and 0.99, with a mean of 0.97 and 0.98 during the first and second trail respectfully. These consistently high scores across folds suggest that the model generalizes strongly to unseen data and is not significantly overfitting. Instead, the results reflect the model's robustness and strong discriminative power for stroke prediction.

Observation and Assumption on Misclassification

```
gender: Most misclassified value = Female (66 cases)
hypertension: Most misclassified value = No Hypertension (82 cases)
heart_disease: Most misclassified value = No Heart Disease (87 cases)
ever_married: Most misclassified value = Yes (81 cases)
work_type: Most misclassified value = Private (57 cases)
Residence_type: Most misclassified value = Rural (49 cases)
smoking_status: Most misclassified value = never smoked (40 cases)
age: Most misclassified range = (65.928, 82.0] (42 cases)
avg_glucose_level: Most misclassified range = (59.88, 100.779] (51 cases)
bmi: Most misclassified range = (24.7, 32.1] (53 cases)
```

Figure 21: Showing misclassification for Trail 1 Random Forest

For trail 1 (no outliers removed), misclassified cases in the Random Forest model show that certain demographic and health-related attributes were more prone to errors.

Female cases were most frequently misclassified, along with individuals who reported no hypertension and no heart disease, indicating the model's difficulty in distinguishing between genuinely low-risk individuals and those incorrectly predicted. Married individuals and those employed in the private sector were also common among misclassified cases, as were rural residents and people who had never smoked.

For numerical variables, the model most often misclassified individuals within the age range of 65 to 82 years, those with average glucose levels between 59.9 and 100.8, and those with BMI values between 24.7 and 32.1. Overall, the findings suggest that the model struggles with cases that outwardly appear lower risk but belong to the positive class, as well as with individuals in mid-to-high ranges of age, glucose level, and BMI where classification uncertainty is higher.

Assumption:

The model struggles with older, relatively healthy individuals whose features are less distinct, and demographic factors like gender, marital status, work type, and residence may contribute to these errors.

Limitations

A major challenge encountered was the imbalance in the dataset, which initially resulted in low model accuracy. Ideally, obtaining more real-world data would address this issue, but given project constraints, the dataset was balanced using SMOTE to generate synthetic samples for the minority class, "strokes." Additionally, there was only a single sample labeled as "other" gender. To avoid complicating the model and the risk of overfitting due to this disproportionately small class, this sample was deliberately removed.

Conclusion

Among the three models, Random Forest clearly delivered the strongest performance especially when outliers were removed, achieving the highest accuracy (94%), the best balance between precision and recall, and an outstanding ROC AUC of 0.98, indicating excellent discrimination between stroke and non-stroke cases. While Logistic Regression and Decision Tree models both performed reasonably well with accuracy around 79-86% and ROC AUC of 0.85-0.96, they misclassified more cases compared to Random Forest. Logistic Regression remains valuable for its interpretability, making it suitable when model transparency is required, whereas Decision Trees offer a simple but slightly stronger alternative. Overall, Random Forest stands out as the

most reliable model for stroke prediction, particularly due to its superior ability to correctly identify stroke cases with high recall, which is crucial in a medical context.

Though the model performed better when removing outlier, caution needs to be taken when considering the context of the data. The dataset is that of medical records, assuming they were correctly recorded, we may be removing edge cases that will affect the model predictability in the future. It may have worked for the vast majority of normal cases but when it come to uncertainty in health of human beings, the model may struggle.

References

Github Project: maryamnasir2000/DSTI-ML-PY-STROKES

Random forests - An ensemble of decision trees | Towards Data Science

Does the risk of stroke from common risk factors change as people age? | ScienceDaily

Heart attack: Younger adults with obesity for decade have higher risk

Diabetes and Cardiovascular Disease in Older Adults: Current Status and Future Directions | Diabetes | American Diabetes Association

What is SMOTE & How Does It Work? - ML Journey

Interquartile range - Wikipedia